Финальный проект

Виталий Косинов
03 марта 2023 г

Выполненые шаги

1.Загрузка данных
- Проверка, что признак равен нулю
- Преобразуем buy_time к дате.

2.Краткий разведочный анализ данных
- Проверка на пропуски
- Распределения
  - Распределение целевой переменной target
  - Распределение признака vas_id
  - Диапазоны дат трейна и теста
  - Распределение id абонентов
- Поиск дубликатов id

3.Сохраняем в pkl

4.Загружаем из pkl

5.Исследование повторяющихся id

6. Объединение merge_asof nearest
   - Модель LAMA
   - LAMA + stacking

7.Объединение merge_asof backward
   - Модель LAMA

8.Объединение merge_asof forward
   - Модель LAMA

9.Исследование дублирования 'id’

10.XGBoost
   - hold out
   - Cross-Validation
   - Бустинг `dart`

11.Обучение модели на всём датасете для прогноза

12.Формирование индивидуальных предсказаний для абонентов

Выбор модели

Опробованы модели:

- На основе фреймворка LightAutoML от Sber:
  - 'linear_l2’ – линейная с L2 регуляризацией (ridge)
  - 'lgb’ – LightGBM с гипер-параметрами по умолчанию
  - 'lgb_tuned’ – LightGBM с гипер-параметрами оптимизированными Optuna
  - Стэкинг: ridge + LightGBM + CatBoost состеканные с LightGBM (Optuna) + CatBoost

Опробованы варианты объединения датафреймов с помощью merge_asof с вариантами:

- nearest
- backward
- Forward

Опробован XGBoost с типами бустинга:

- gbtree (по умолчанию)
- dart (не склонный к переобучению)

Выбрана модель XGBoost как показавшая наивысшее качество.

Предложение для абонентов

Для составления предложений для абонентов был использован следующий принцип:

1. Модель обучается на всей обучающей выборке.
2. Модели подаются на вход каждый класс подключённой услуги
3. По каждому варианту услуги считается вероятность подключения
4. Выбирается услуга, имеющая наибольшую вероятность подключения

Алгоритм можно улучшить, установив порог вероятности для рекомендации услуги. Таким образом, если услуга с максимальной вероятностью для данного абонента тем не менее, низка (порог необходимо подбирать ориентируюсь на потребности бизнеса), услугу не стоит подключать. Это позволит избежать негативного эффекта снижения лояльности клиентов от «несработавшей» рекомендации.

Ссылки:

Проект на GitHub